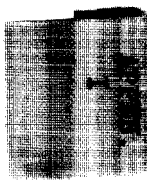
Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6



Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

PROBLEM OF MACHINE TECHNIQUE IN SCIENTIFIC INFORMATION

BY L. I. GUTENMAKHER

Vestnik Akademii Nauk SSSR, No. 8, 1952, pp. 46-52

(Original in Battelle Memorial Institute Library)

The problem of the development of a machine technique in scientific information, set up by the President of the Academy of Sciences of the USSR A. N. Nesmeianov, represents a basic complex problem, the solution of which will result in a radical change in scientific-technical methods of information, will contribute to the better organization of scientific research and will greatly foster the application of scientific advancement into the national economy.

The steady growth of the number of scientific research works and technical developments has resulted in an avalanche of printed material which completely entangles scientists and practicing engineers.

The total amount of printed material produced by the human race has to be calculated in the range of hundreds of millions. At the present rate of research, the amount of material is such that the library holdings will double in 10 to 15 years. After 50 to 60 years, it may be expected that library holdings will increase 16 to 20 times. Specialists in some fields of science are not able to follow the progress in adjoining fields of science and technique. For the purpose of systematization and selection of bibliographical data and compiling bibliographies, a large army of bibliographers is at work to help out scientists.

Contemporary practice has required the solution of complex technical and scientific problems in the shortest possible time, taking as a basis the existing data. Most of the scientists' time is used for selecting literature and obtaining exhaustive information.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Because of the great amount of information scattered among large numbers of magazines, books, and symposia, the process of finding the necessary information takes such a long time that it is easier to conduct an experiment and to make a calculation than to find the description of such in the literature.

Attempts at classifying the information material by using different library-classification methods could not lead to the radical solution of the problem.

The classification as a method of organized arrangement of material is characterized by the selection of certain criteria in the form of a basis for dividing information into individual "noninterseeting" groups. For example, bacteria are classified according to their morphology, their pathological action, the conditions of their life and growth, etc. Chemical compounds are classified according to their composition, chemical structure, physical properties, and according to the field of their application. Electronic appliances are classified according to fields of application, capacity, characteristics, etc.

The classification consists basically in a variation of some type of criterion. It is impossible to use all possible combinations of all criteria.

An attempt to create such a comprehensive and ramified classification so that in its final stage it will produce the simplest possible answer was doomed to be unsuccessful, too.

The theory of combinations makes it possible to determine easily the "astronomical" numbers which are obtained by computing the possible number of elementary questions in such a system. Such a number even in the case of a very small amount of initial data will have at a base of 10 a power

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

of the order of 1000 (10^{1000}).

In connection with the above difficulties, the selection and perusing of literature has to be in charge of individual scientists and specialists. In general, bibliographers and special sections of libraries, ministories, institutes, and universities are helping out in the bibliographical research of interested persons.

For example, let us attempt to calculate the amount of work necessary to satisfy the primary needs of scientific information.

There exists in the Soviet Union several millions of engineers, technicians, scientific workers, etc.

Let us assume that each of them required only one time a year some scientific information (bibliography in a specific field). It will amount to around 3 to 6 millions of requests each year or 10 to 20 thousand requests a day.

Further, we shall assume that the selection of individual information is made from material, the volume of which is, for example, represented by 1000 pages of text. We shall assume also that one person per day may thoroughly scrutinize 100 pages of such text. In such a case, each information will require an average of 10 man-days.

Therefore, for performing the work of requested information, it will be necessary to have 100,000 to 200,000 qualified researchers scrutinising available material.

Our approach is based not on the existing solution concerning the research work, but on the desirable scale of such.

Even if the volume of desired information would be lowered to a great extent, a very great amount of work for procuring information necessary in scientific research still would be required.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

The constantly growing scientific level of Soviet specialists requires the technique of procuring information which will ensure for them an average of not only one piece of information per year, but several.

It may be assumed that because of insufficient information, part of the efforts and means of scientific institutions are used purposelessly for duplicating already performed investigations.

Considerable time is being spent for the selection of information as a first step of any large research, because, before initiating such scientific research, it is necessary to become familiar with the data from all literary sources. Quoting V. V. Maiakovsky, it is possible to say that a scientist selecting scientific-information material, investigates "thousands of tons of 'verbal' ore".

Information material accumulated in libraries represents a tremendous potential richness, having more advantages the better the scientific information service is organized.

I. V. Stalin indicates that the mechanization of labor represents the force without which it is impossible to withstand the impact of the new scales of production. This indication could be fully applied to the problem of scientific information.

In any case, the mechanization is not limited to a simple increase in the rate of material selection. The search for information on problem A may be located together with information on problem B; it may happen that information on problem B will be required. Therefore, it may be advisable to establish the relation between A and B and to determine the type of such relationship.

The possibility of a rapid extraction of compiled data on individual problems will result in addition to a series of other benefits, the decrease

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

and possibly the elimination of steadily increasing "bureaucracy" in the science. At the present time, specialists even in connected fields hardly understand each other. More and more labor is being spent for finding and applying analogies in processes, phenomena, and structures existing in different fields of sciences. The preparation of material for information-bibliographical machines will require the generalization of most diverse research and investigations.

According to the idea of Academician A. N. Nesmeianov, it is necessary to create the possibility of perusing the content of information by means of a machine, in relation to a given problem, taking as a basis the identification of such problems with a series of independent criteria.

In such cases, the content of each individual research must be identified in the research report by a certain number of simple elementary phrases, the sizes, facts, statements, criteria. Scientific hypotheses, ideas, results of experiments, principles of devices, specific physical constants, time, location, and other information data must be presented in a condensed and well-defined form.

In first approximation, it may be assumed that an average scientific paper will contain between 100 and 200 such sentences. The analysis of all incoming material, formulation, and registration of elementary sentences may be performed according to the rules and instructions established by authors of articles themselves or by specially provided personnel. The major part of such work could be performed by the already established Institute for Scientific Information during the period of waiting abstracts on work in different fields of science and engineering.

The selection of accumulated material according to a given problem should be performed by machine.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

The problems themselves must be formulated as simply as possible, in the form of elementary sentences. Simultaneously, a large number of such problems could be included in the question. Only during checking the obtained information would it be necessary to consider the information related to all problems simultaneously.

The answer (solution) must contain the enumeration of selected information (selected bibliography) in the form of serial numbers of work registered in special libraries and indicate its content.

The problem of obtaining Photostats or originals of the selected work must be solved separately.

This problem also may be solved using mechanized devices (automatic cameras and other devices).

The basic problem of the automatic device (for brevity, we will call this device a machine) is to compile bibliography according to the combination of a series of criteria (problems or questions). At a large number of such criteria, the number of possible combinations, practically, is unlimited.

In certain cases, undoubtedly, a combination may be required which is not reflected in any available reference. A negative answer to such a question is also useful, because it will indicate the newness of the problem or of the investigation or its development.

Whereas the selection is performed according to the content of the machine, according to the basic idea, may answer most [illegible] combination of given criteria.

For example, information on physical constants of molecules is selected and arranged. The problems to be solved by the machine will be formulated as follows: To find on the basis of existing bibliographical data the identification number of works in which such constants of molecules

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

of certain compounds possess values in the range established by the problem posed. In such cases, the number of criteria may be different, ranging from one to the maximum number possible, depending on the stated problem.

However, the possibilities of the machine built according to the indicated principle are much greater. In a series of cases, the person requiring the solution of the problem may be interested not in the bibliography as such, but in the analysis of the content of the articles.

For example, information material on chemical kinetics and on data concerning the mechanism of chemical reactions is selected and arranged correspondingly to the stated problem.

Problems confronting the machine selection of information in such cases may be formulated as follows:

1. To find articles in which slow reactions are discussed, that is, such reactions in which a "pre-exponential" term has a value within certain given limits.

2. To find articles in which it is indicated that a certain reaction proceeds at a definite rate indicated in the problem, and to indicate data concerning the temperature and concentration.

Therefore, it is possible, on the basis of given criteria, to require information of all (or part) of the data numerically related to such criteria. In the beginning, it may appear that limitation of the solution by certain numbers of criteria without indicating the source of information has no meaning. But if the tremendous amount of material that may be examined using the machine method is considered, the expediency and effectiveness of such machine method will become apparent.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Many times it is necessary to require very urgent answers concerning the relation of certain values, concerning the coincidence or noncoincidence of a series of signs, etc. For example, in the above-indicated case, it is interesting to check whether there are contradictory data for corresponding reactions under the very same conditions or under different ones. (For example, a reaction in the temperature range from t_1 to t_2 follows the general kinetic equation; however, the temperature range from t_2 to t_3 indicates the presence of a more complex chain mechanism.)

The machine method makes it possible to extract very rapidly and thoroughly accumulated data, to compare different factors, to analyze data etc.

However, to realize such a machine system, it is necessary to solve a series of quite complex problems.

1. CREATING AN ECONOMICAL, WELL-DEFINED SYSTEM FOR RECORDING INFORMATION

The science that produced the above problem prepared the means for its solution. It is sufficient, for example, to recall the existence of chemical formulas indicating the structure of substances. The theory of [illegible] presents the possibility of expressing different physical values [illegible] a small number of basic values. Thus, to characterize phenomena [illegible] in mechanics, it is possible to assume length as L, mass as M, [illegible] T as basic values, and the dimension of the mechanical [illegible] expressed as L^2MT^{-2}; speed as LT^{-2}; density as [illegible], power as L^2MT^{-3}; etc.

For the characteristics of electromagnetic phenomena, a fourth value is added to these basic values - dielectric permeability ε or magnetic permeability μ.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Therefore the words "electric field intensity" may be written by the dimension formula: $L^{-\frac{1}{2}} M^{\frac{1}{2}} T^{-1} \varepsilon^{\frac{1}{2}}$. If in the text of information the words "electromotive force", "potential", "intensity", "voltage" appear, all of them may be expressed by the formula: $L^{\frac{1}{2}} M^{\frac{1}{2}} T^{-1} \varepsilon^{-\frac{1}{2}}$.

A compilation of a particular dictionary, generalising the [illegible]fold specialized dictionaries now existing will give the possibility of finding information in the most unexpected places.

Many examples are known where "new discoveries" in one field of science have been used for a long time. For example, a feedback in a mechanical booster of a regulator of a steam engine has been used for increasing stability of operation for about 80 years, but for electronic amplifiers, it was "rediscovered" only in the 30's of the present century and many years after this for magnetic amplifiers.

The trend to a greater generalisation exists in every field of science. Considerable achievements have been attained in this direction by Soviet scientists, e.g., in the field of the theory of oscillation. To generalise the available material, the experience of the theory of similarity and analogies of physical phenomena may be useful.

The method of analogy is based on V. I. Lenin's postulate: "The unity of nature is indicated in an astonishing analogy in differential [illegible] in different fields of science". Certain mathematical [illegible] electrical, mechanical, hydraulic, and other phenomena are well known. Academician A. N. Krylov indicated, that such "analogies between problems of completely different fields, but resulting in similar differential equations, are existing in large numbers".

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

What similarity may exist between the calculation of the motion of planets governed by their own gravity and that of the sun and rolling of the boat, or between the determination of so-called century inequality in the motion of planets and rotary vibration of the diesel multicylinder engine crankshaft when operating a ship propeller or electrogenerator? However, if such a formula and equations were described without words, then it would be impossible to determine which of these problems is being solved, the equations used being exactly the same.

Therefore, the presence of analogy in such various phenomena make it possible to describe them in the following form:

a) by a system of generalised equations

b) by formulas of dimension of existing values.

c) by dimensionless values (criteria of similarity).

d) by a series of elementary sentences, indicating the purpose and results of investigation or development.

The most exact and laconic formulation in many cases may be obtained by using mathematical formulas. This symbolic, economical form of recording different concepts is changing and is in the process of further development.

For example, the recording of an algebraic [illegible] [illegible] years ago would be represented like this:

[illegible] planes [illegible]

At the present time, [illegible] of very complex logical concepts, operations, and [illegible] algebra of logics) has been created.

The utilization of the arsenal of mathematical means produces very commendable results. Experience from the theory of similarity and the

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

introduction of dimensionless values (criteria of similarity) for the evaluation of encountered values in relation to basic units also may bring considerable help in solving problems.

One should remember the great experience in the development of very exact and clear formulations of statements in patent practice. As is well known, the formula of invention should be recorded in the form of a series of separate elementary sentences, in which case each sentence must be short and self-containing.

If all of these efforts in different fields are summarized, the obtained result will have a great individual scientific importance.

The development of the technique of scientific information according to the idea of Academician A. N. Nesmeianov, requires the creation of a basis for generalizing information which logically is the next stage of development of the theory of similarity and of the unity of phenomena.

The successful solution of this problem will result in even greater effective utilization of the idea of dialectic materialism concerning the manifestation of the unity of nature in the development of science and [illegible] the general bonds and interrelationships between different [illegible]

THE DEVELOPMENT OF A RATIONAL SYSTEM OF CLASSIFICATION OF MATERIAL

[illegible] the system of recording [illegible] must be used for, the system of classification [illegible] to find the material.

[illegible] machine method would make it possible to peruse all accumulated material within a certain time, for example, 10 to 20 minutes, then the problem of setting up a system of classification will disappear by itself.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

However, the amount of information data is so great that even at highest speed of operation it would be impossible to peruse available material in 10 to 20 minutes.

The preliminary calculations showed that if in one year 100,000 abstracts, each of them containing 100 elementary sentences with 10 words each, are received, it will amount to 100,000,000 words of text per year. For perusing such an amount in 10 minutes, about one hundred thousandth part of a minute for each word will be required.

When material that has accumulated for 10 years will have to be perused, then a speed of operation must be increased up to a millionth part of a second or the time of operation lengthened.

Undoubtedly, in many important cases, [illegible] avoid omitting any material, an increase of the time of perusing all available material according to the given criteria may be recommended.

However, in most cases, the [illegible] could be limited. For example, if information on [illegible] required, it would be foolish to peruse the [illegible]. Therefore, it would be of advantage to establish [illegible] system of classification to avoid loss of time in the search.

In connection with the above, it would be advisable to develope an [illegible] period of time and later on to [illegible] information (method of [illegible]

In machine [illegible] the commutators must be constructed [illegible] could be easily rearranged when the method of classification is changed.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6

Because the machine method is a high-speed device for research, the system of classification must be simple. The structure scheme of classification will probably have the form of a "tree of knowledge" with a different number of "branches" in sections.

During the setting up of the program of information research, it would be possible to record a considerable number of "addresses" of such "branches" in which the presence of available material may be expected.

The modern technique of commutation makes it possible to establish for each division of information not one, but several parallel contacts.

If, for example, it is known that given information simultaneously presents interest for chemistry, physics, biology, and for measuring technique, then the "addresses" of all divisions of the fields of science connected by general interest may be ascribed to this information.

We are not considering the engineering part of this problem. The difficulties here seem to be even greater, but modern engineering is able to cope with them.

The development of such a machine technique of information research is of considerable scientific and engineering importance, because the obtained results could find application in automatic [illegible], telemechanics, and communication. The [illegible] presented seems to [illegible] prospects. [illegible] its importance [illegible] character, this [illegible] handled [illegible] Academy of Sciences [illegible].

[illegible]/tms/wkf

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150006-6